Rapid Assessment of Hydrocarbon Contamination in Soils Using Mid-IR Spectroscopy & Binary Classifiers

Deeksha Beniwal*, Sean Manning, Georgios Tsiminis

SETAC 2023 Conference, Townsville
Session: Emerging Monitoring Technologies & Developments in Environmental Chemistry and Ecotoxicology

Ziltek

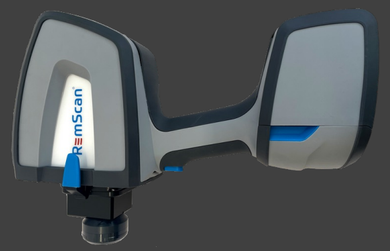
Motivation

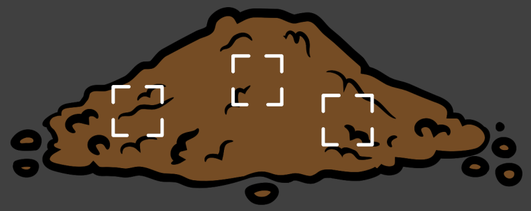
- Rapid on-site assessment is needed for efficient remediation.
- Traditional lab tests are expensive, slow and resource-intensive.
- RemScan is a fast, cost-effective measurement solution.
- Extensive work done to calibrate the instrument.
- New calibration method being developed for Remscan.
- Improve speed and accuracy of measurements.

Ziltek

In-field operation

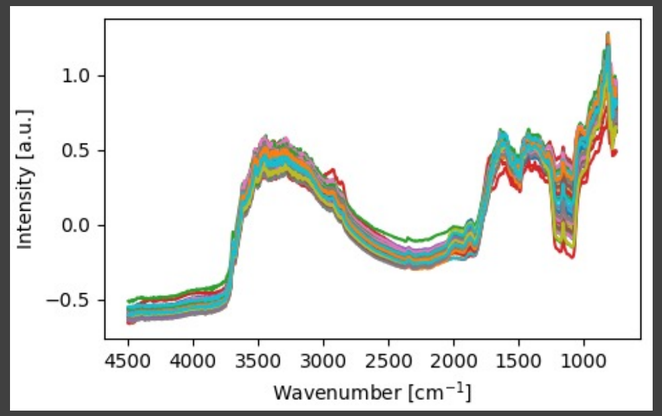
RemScan





Contaminated Soil

Mid-Infrared Spectra



Binary Classifier

CONTAMINATED

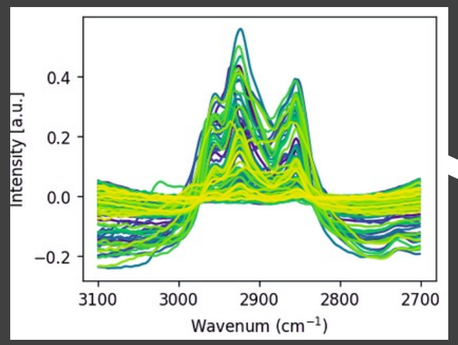
CLEAN

*Common industry threshold = 1000 ppm/ 0.1% contamination

Ziltek

Methodology

TRAINING SPECTRA



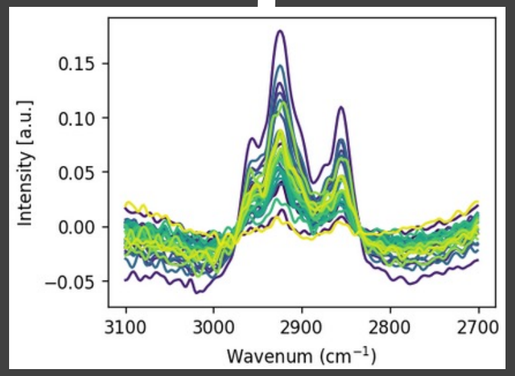
LABELS

	filename	tph	class_id
0	7832_13_01-29Jan14-1577_0000 (2)	100000	0
1	7832_13_01-29Jan14-1577_0001	100000	0
2	7832_13_01-29Jan14-1577_0002	100000	0
3	7832_13_01-29Jan14-1577_0003	100000	0
4	7832_13_01-29Jan14-1577_0004	100000	0
...	...	...	...
995	17738_12_09-03Feb14-1577_0001	1000	1
996	17738_12_09-03Feb14-1577_0002	1000	1
997	17738_12_09-03Feb14-1577_0003	1000	1
998	17738_12_09-03Feb14-1577_0004	1000	1
999	17738_12_10-03Feb14-1577_0000	781	1

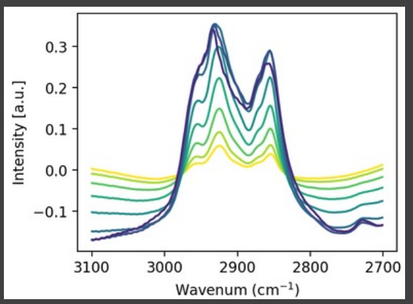
TRAINING MODEL

PREDICTION

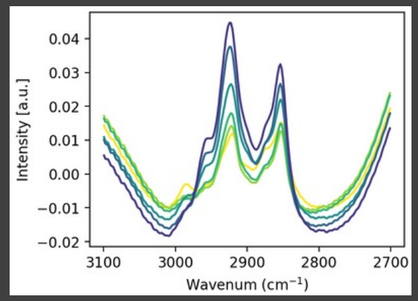
TEST SPECTRA



Contaminated:
TPH>1000 ppm



Clean:
TPH<1000 ppm



Ziltek

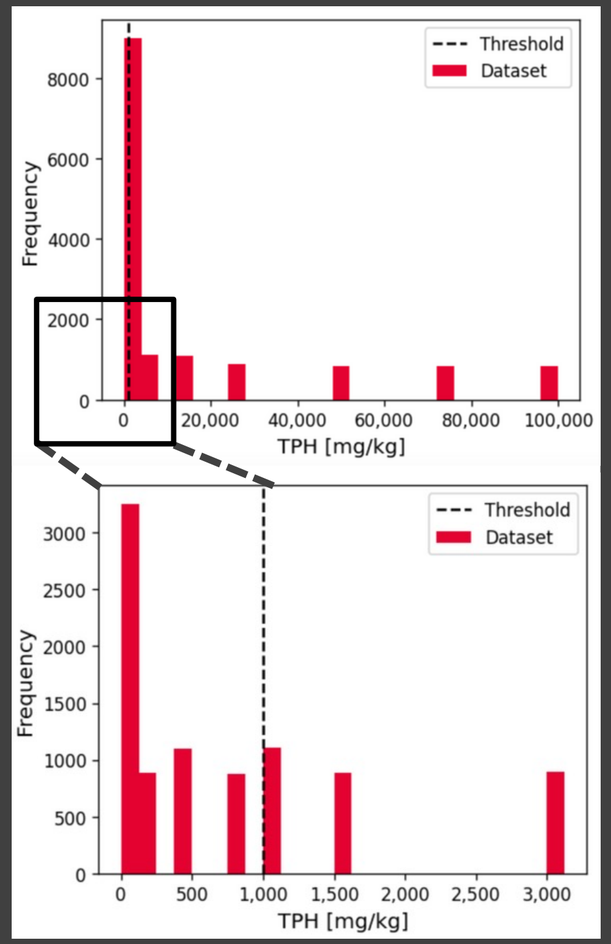
Binary Classifiers

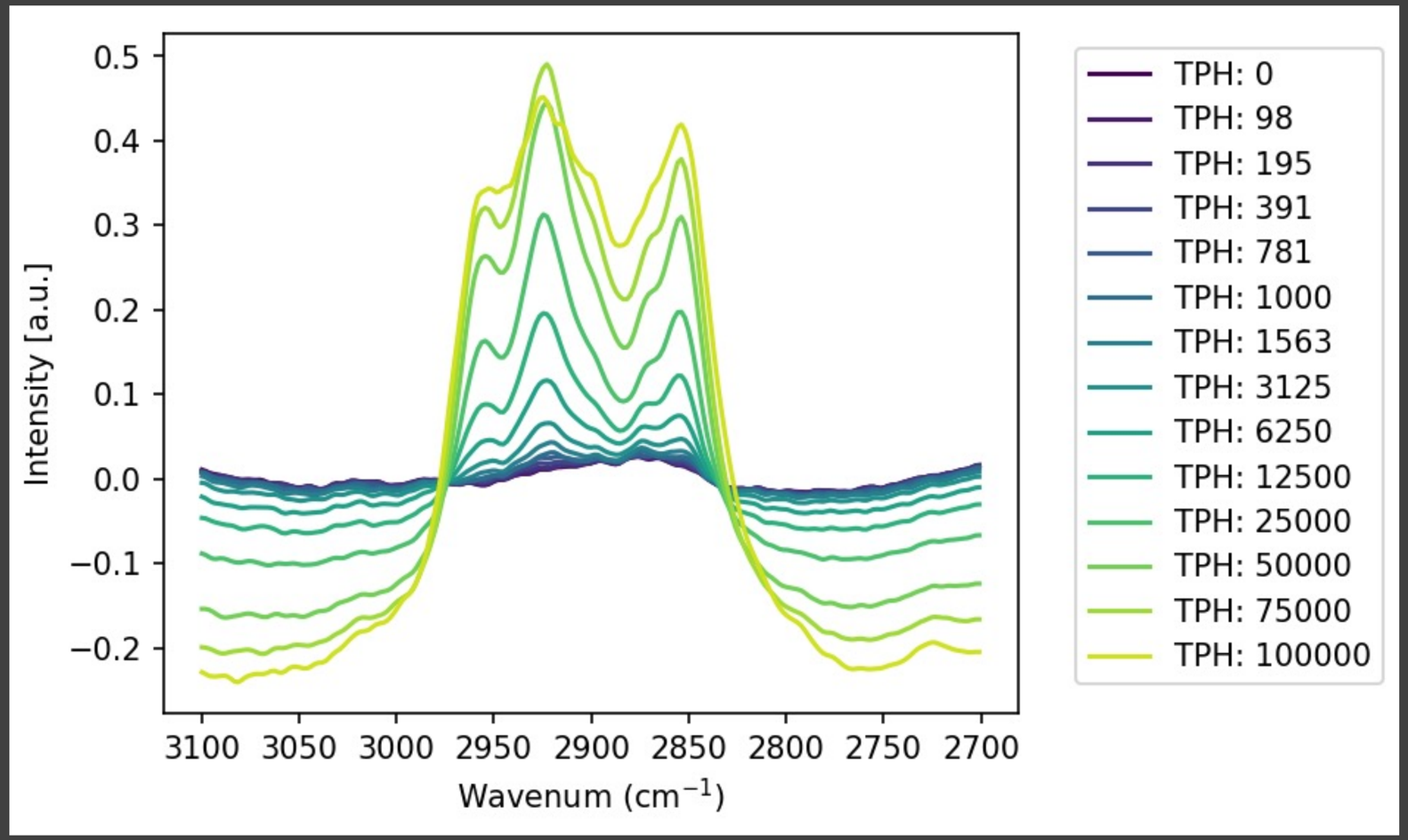
- Logistic Regression
- Support Vector Machine
- K-Nearest Neighbours
- Decision Tree
- Random Forest
- Adaptive Boost
- Gradient Boost

Ziltek

Training Dataset

Total: 14,575

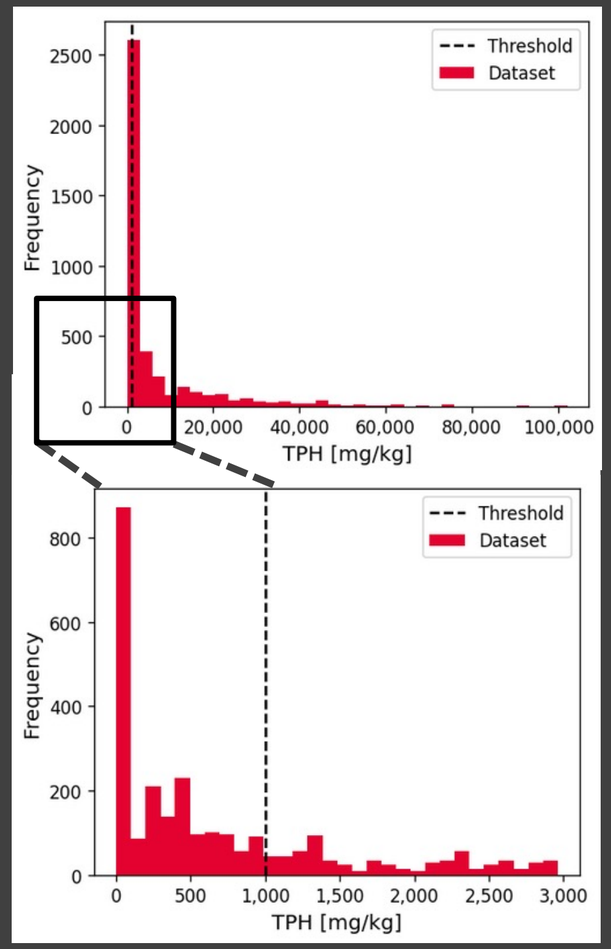


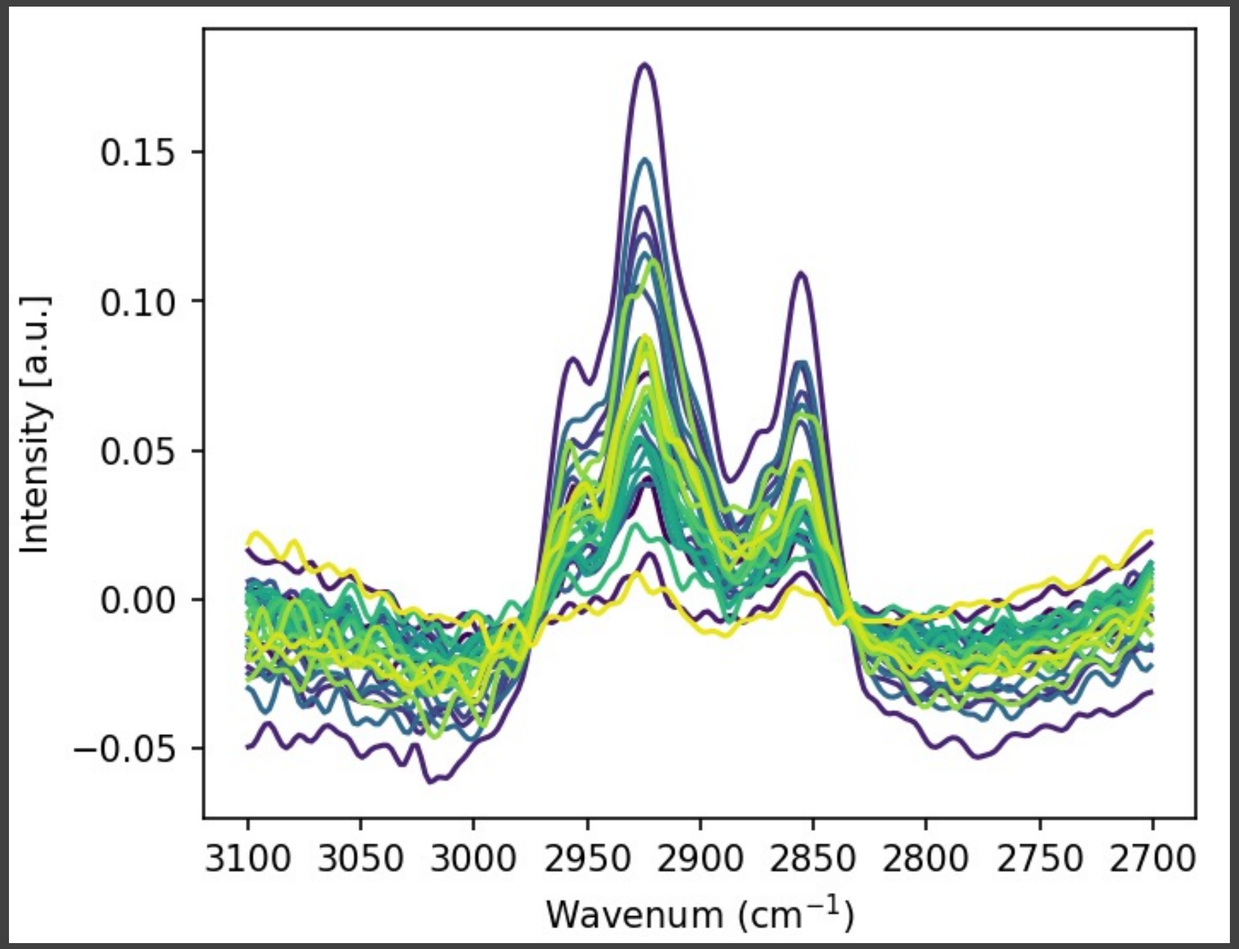


Ziltek

Testing Dataset

Total: 4,045





Ziltek

Performance Metrics

- Accuracy
- Macro F1 score
- Matthew's Correlation Coefficient (MCC)

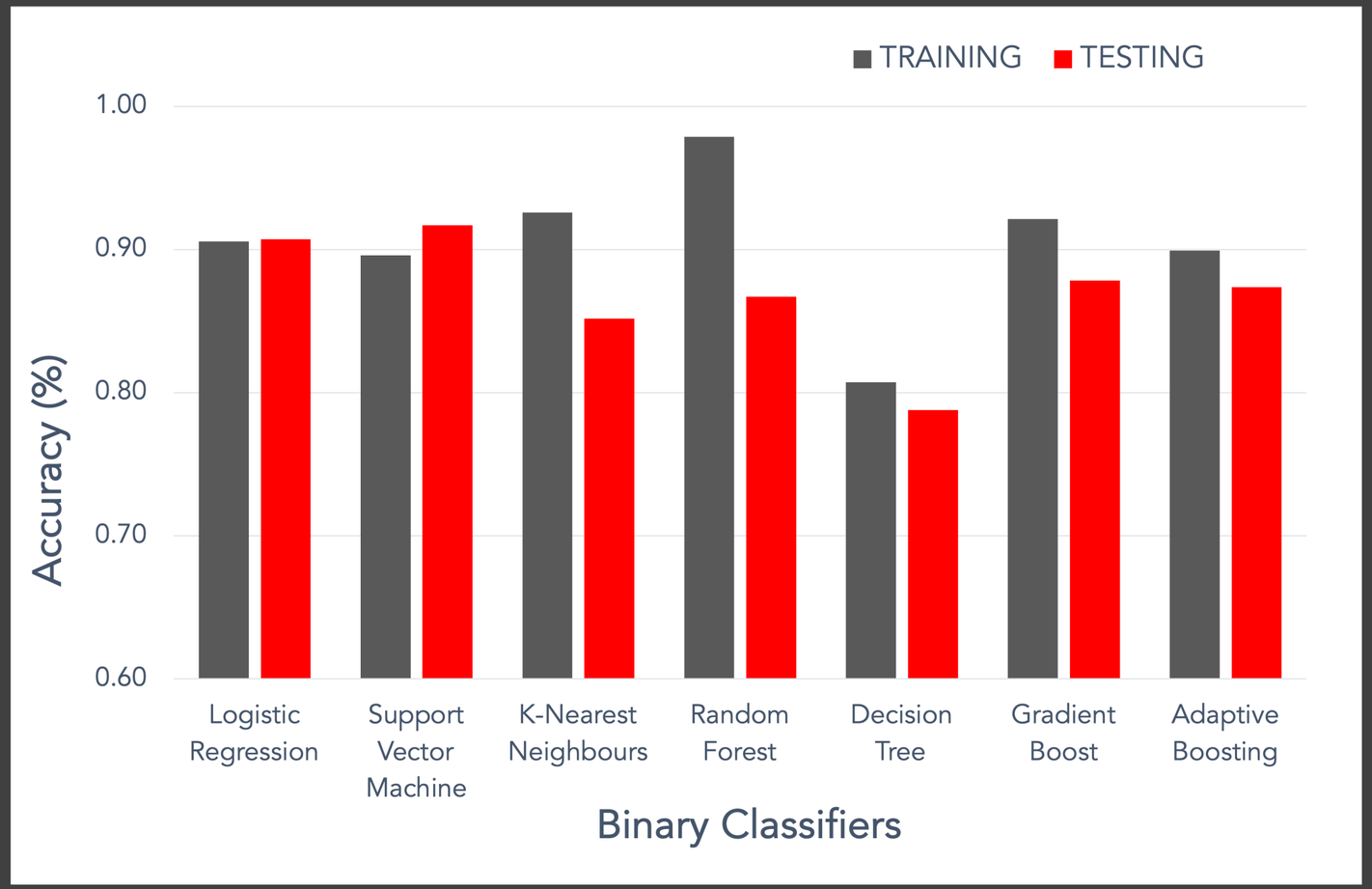
Ziltek

Prelim. Results

CLASSIFIER PERFORMANCE

Optimal classifier [Accuracy]

- Training - Random forest [98%]
- Testing - Support Vector Machine [92%]



Ziltek

Prelim. Results

CLASSIFIER PERFORMANCE

- Same results across all metrics

- Random forest
  - Optimal for training data (98%)
  - Sub-optimal for testing data (87%)
  - Evidence for over-fitting

TRAINING

Classifiers	Accuracy	F1 score	MCC
Logistic Regression	0.91	0.91	0.81
Support Vector Machines	0.90	0.90	0.80
K-Nearest Neighbours	0.93	0.93	0.85
Random Forest	0.98	0.98	0.96
Decision Tree	0.81	0.80	0.65
Gradient Boost	0.92	0.92	0.85
Adaptive Boosting	0.90	0.90	0.80

TESTING

Classifiers	Accuracy	F1 score	MCC
Logistic Regression	0.91	0.91	0.82
Support Vector Machines	0.92	0.92	0.83
K-Nearest Neighbours	0.85	0.85	0.72
Random Forest	0.87	0.87	0.75
Decision Tree	0.79	0.79	0.60
Gradient Boost	0.88	0.88	0.76
Adaptive Boosting	0.87	0.88	0.76

Ziltek

Prelim. Results

CLASSIFIER PERFORMANCE

- Same results across all metrics

- Random forest
  - Optimal for training data (98%)
  - Non-optimal for testing data (87%)
  - Evidence for over-fitting

- Support Vector Machines [SVM]
  - Sub-optimal on training data (90%)
  - Optimal on testing data (92%)

TRAINING

Classifiers	Accuracy	F1 score	MCC
Logistic Regression	0.91	0.91	0.81
Support Vector Machines	0.90	0.90	0.80
K-Nearest Neighbours	0.93	0.93	0.85
Random Forest	0.98	0.98	0.96
Decision Tree	0.81	0.80	0.65
Gradient Boost	0.92	0.92	0.85
Adaptive Boosting	0.90	0.90	0.80

TESTING

Classifiers	Accuracy	F1 score	MCC
Logistic Regression	0.91	0.91	0.82
Support Vector Machines	0.92	0.92	0.83
K-Nearest Neighbours	0.85	0.85	0.72
Random Forest	0.87	0.87	0.75
Decision Tree	0.79	0.79	0.60
Gradient Boost	0.88	0.88	0.76
Adaptive Boosting	0.87	0.88	0.76

Ziltek

Prelim. Results

DATA PREPARATION

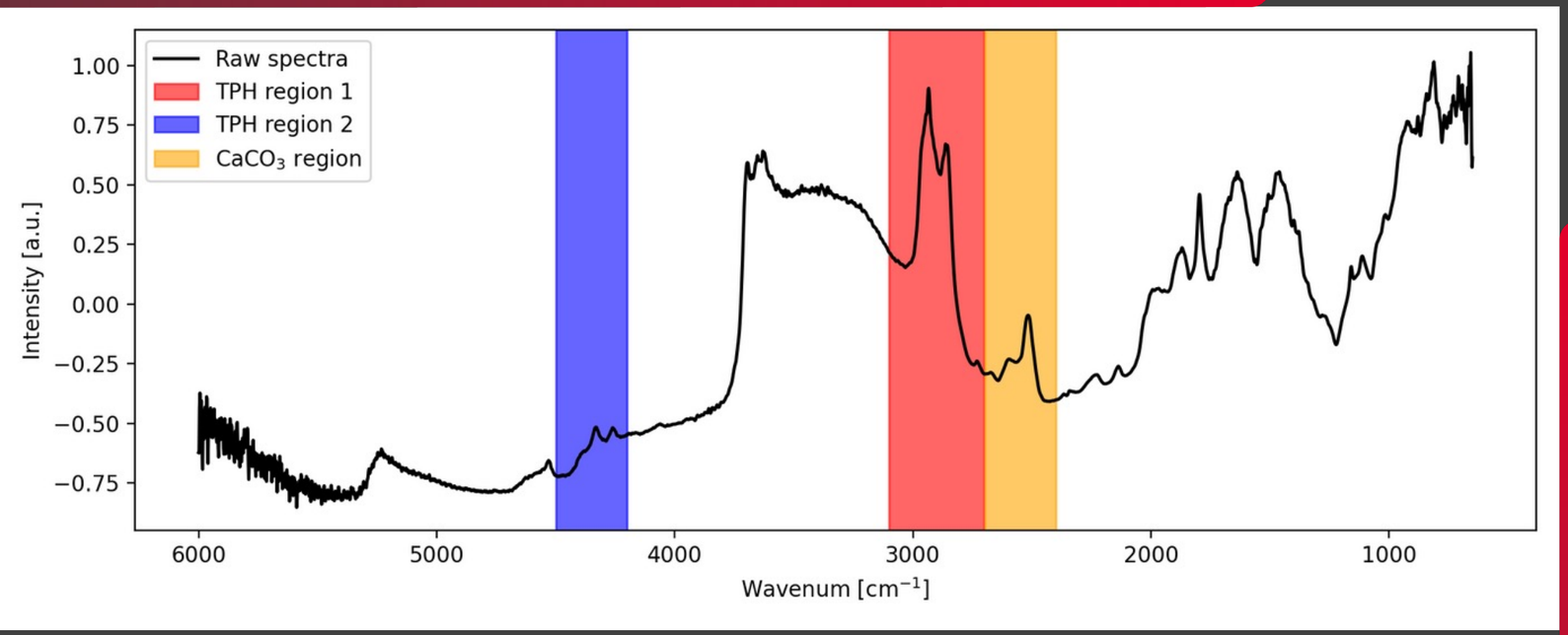


FIG: A typical mid-infrared spectrum of a contaminated sample. The red, blue and orange regions indicate the first TPH region, second TPH region and the calcium carbonate fingerprint region, respectively.

PREPROCESSING

- Tested three scenarios
- Best performance with minimal preprocessing (detrend)

FEATURE SELECTION

- Select spectral regions of interest
- Three different combinations tested
- Best performance with one TPH region
- Computationally efficient

Ziltek

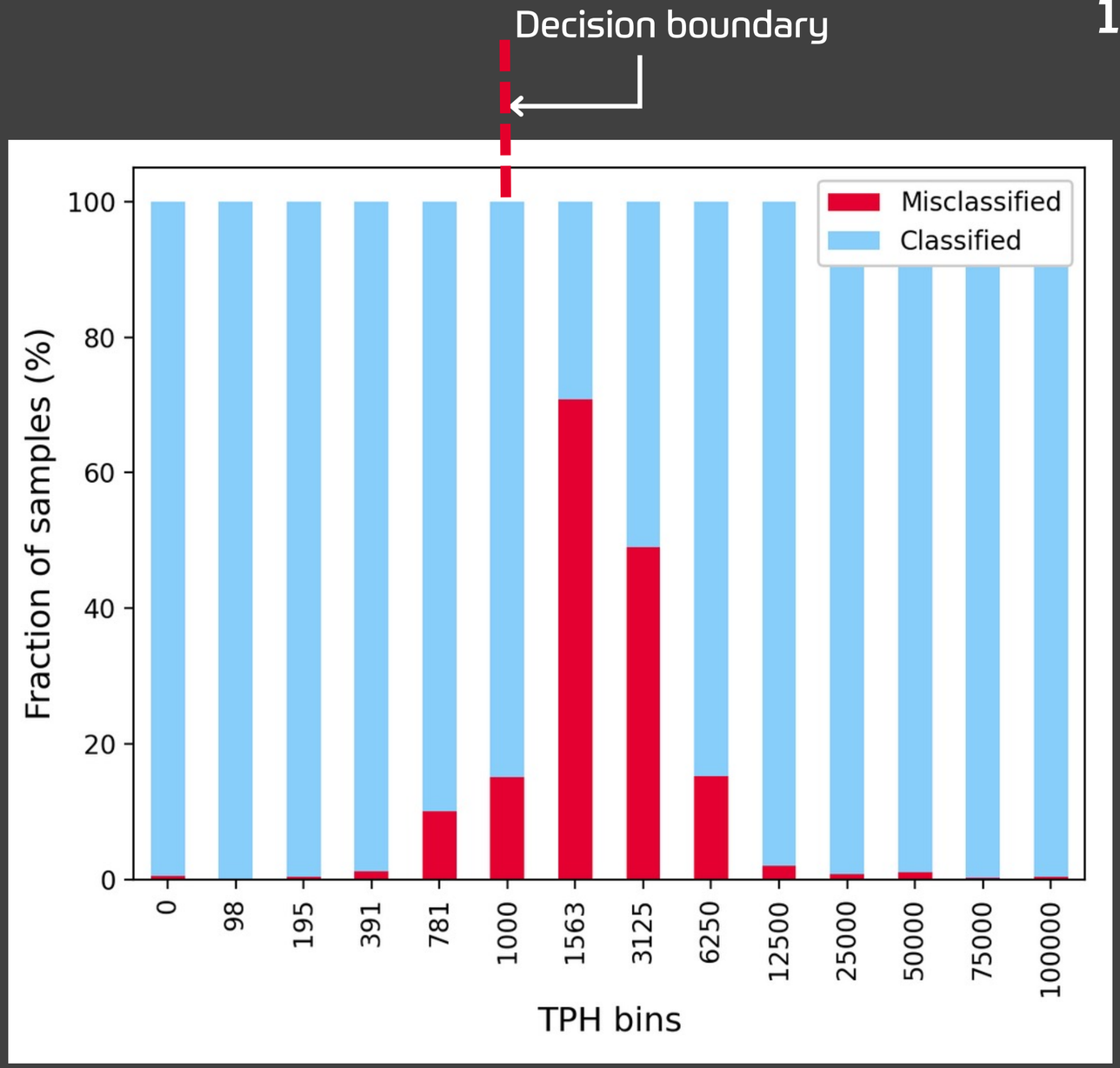
Prelim. Results

PREDICTION DISTRIBUTION

Data-type = Training
Preprocessing = Detrend

Binary Classifier = SVM
Accuracy = 89.2%
F1 score = 0.90
MCC = 0.79

Highest misclassified samples just above threshold



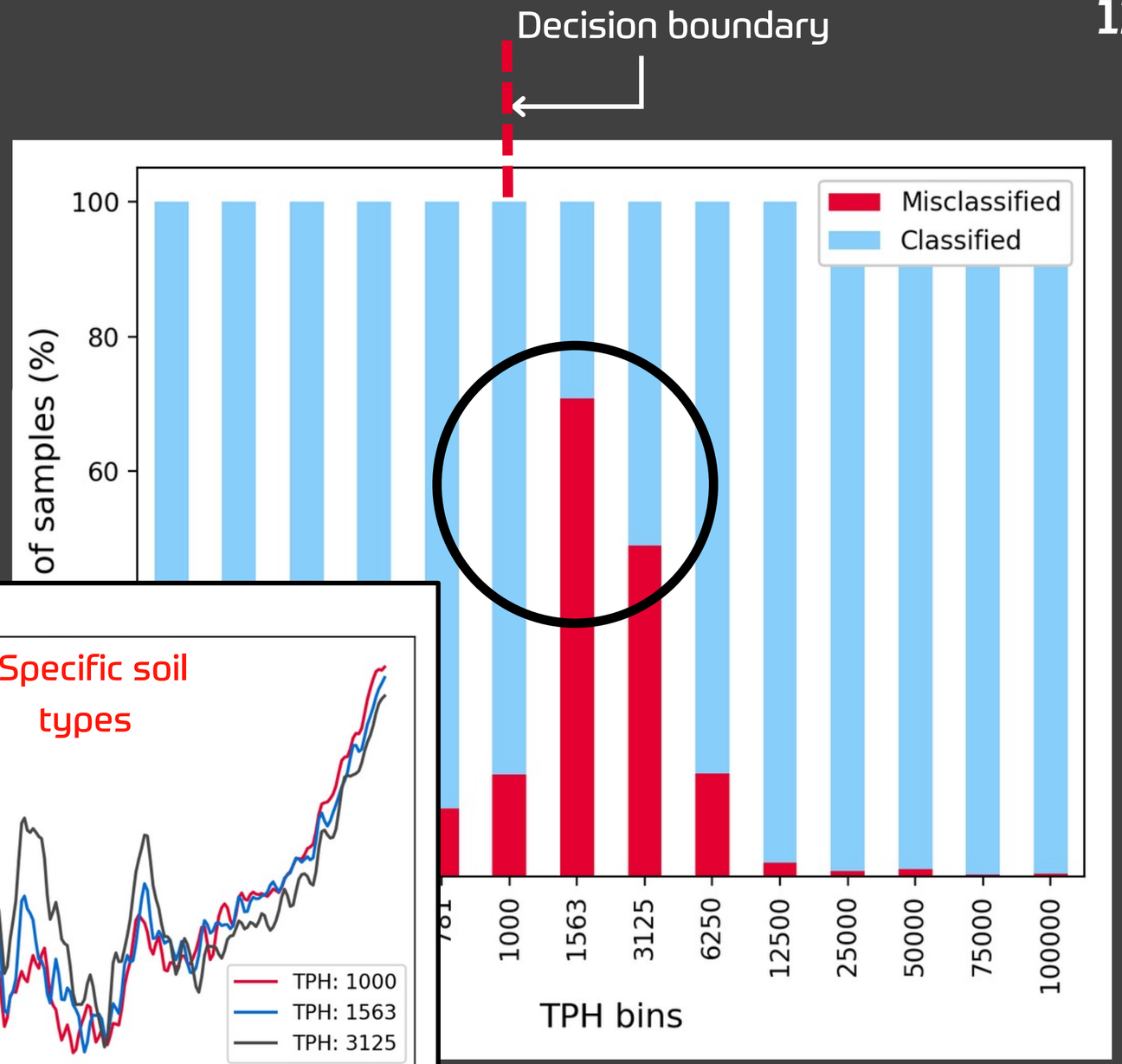
Ziltek

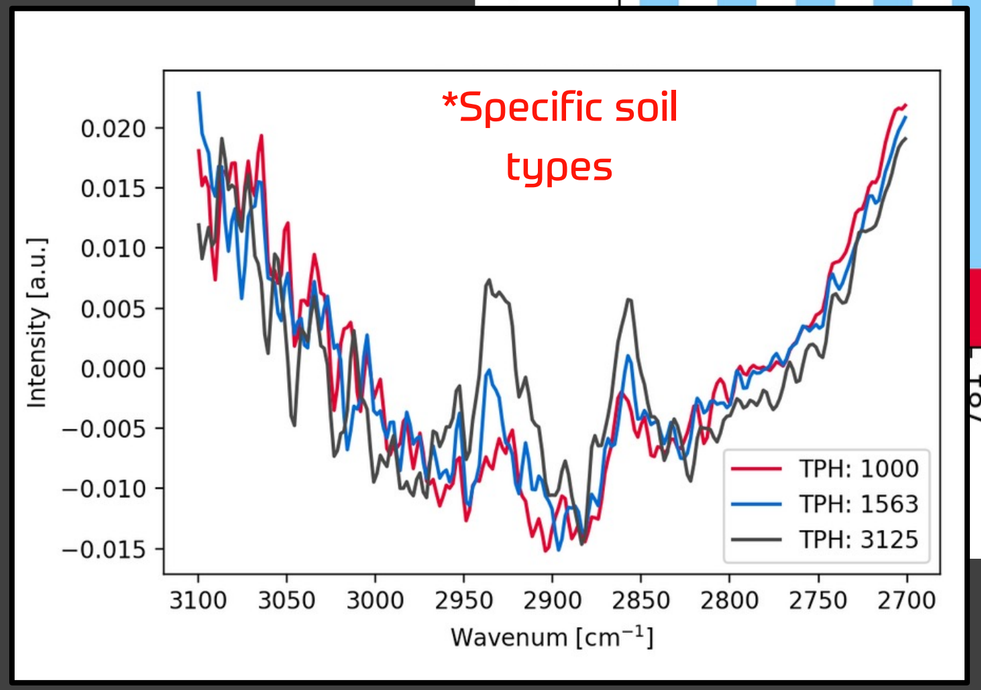
Prelim. Results

PREDICTION DISTRIBUTION

Data-type = Training
Preprocessing = Detrend

Binary Classifier = SVM
Accuracy = 89.2%
F1 score = 0.90
MCC = 0.79





Ziltek

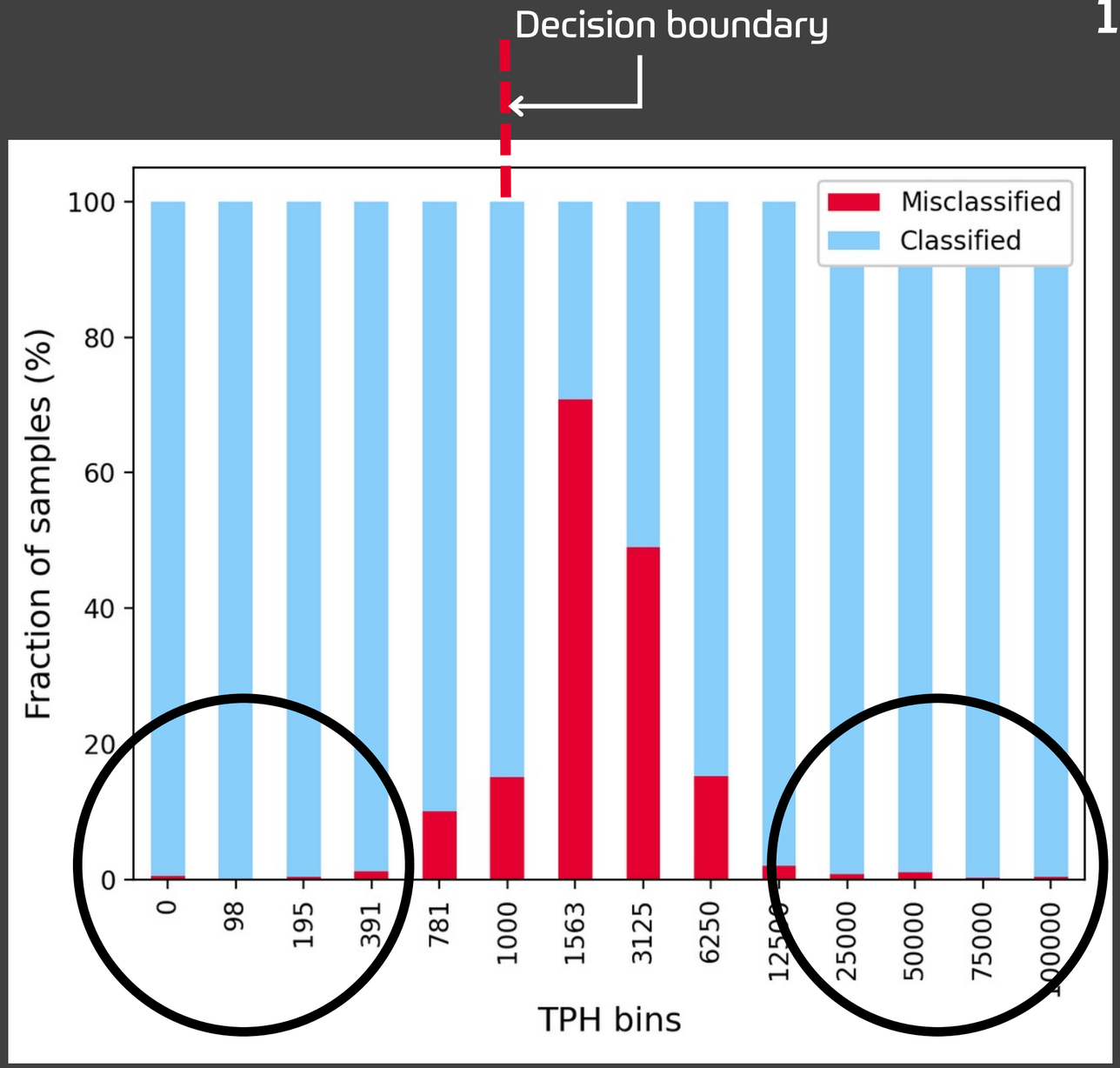
Prelim. Results

PREDICTION DISTRIBUTION

Data-type = Training
Preprocessing = Detrend

Binary Classifier = SVM
Accuracy = 89.2%
F1 score = 0.90
MCC = 0.79

Misclassified samples at Low and High TPH concentrations.



Ziltek

Edge case analysis

Soil Organic Carbon (SOC)

- Known overlap between TPH-sensitive IR peak & natural organic matter

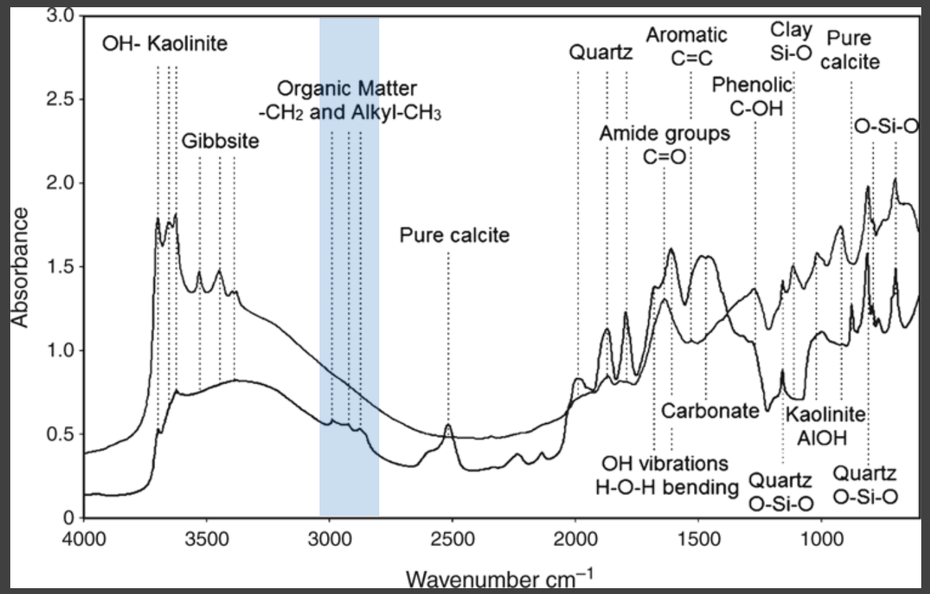
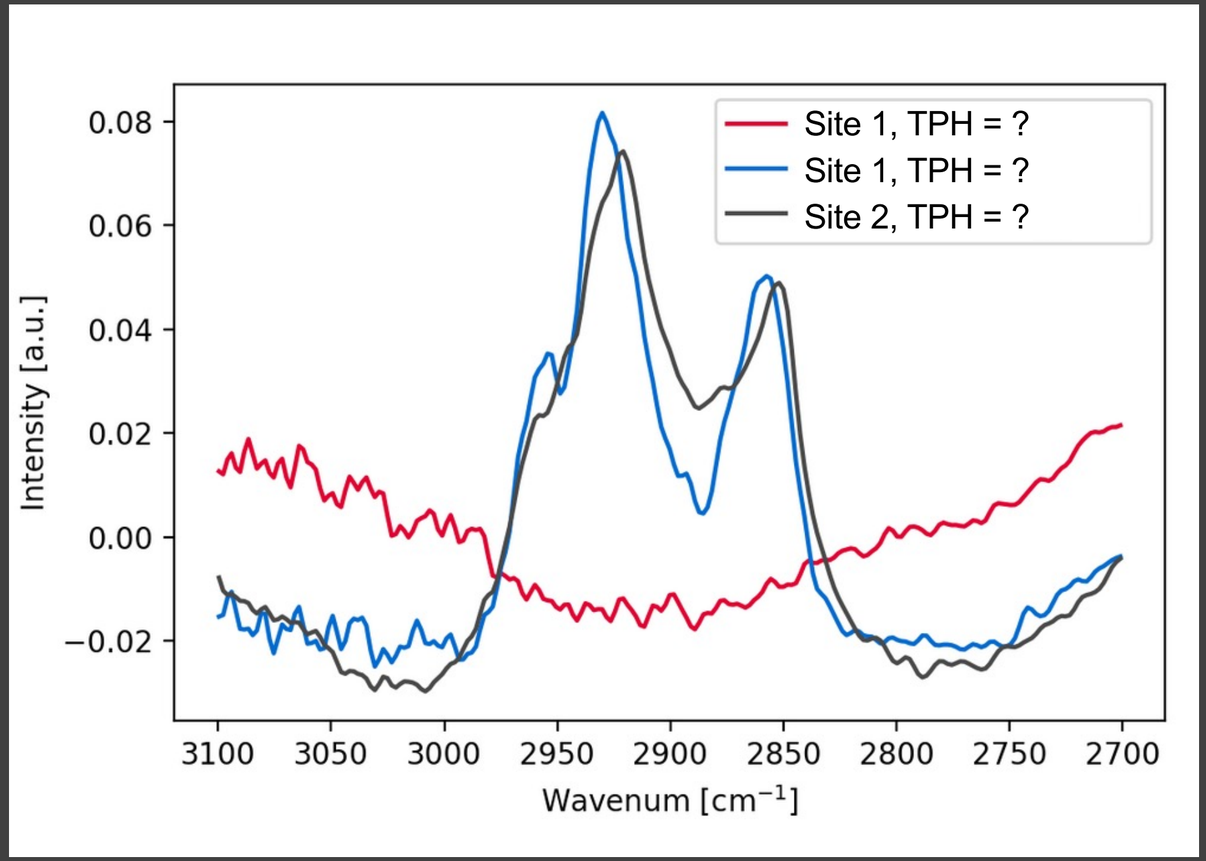


FIG: Representative soil mid-IR spectrum showing absorptions related to the mineral and organic composition of soil.

Reference: F. Le Guillou et al., How does grinding affect the mid-infrared spectra of soil and their multivariate calibrations to texture and organic carbon?. DOI: 10.1071/SR15019



Ziltek

Edge case analysis

Soil Organic Carbon (SOC)

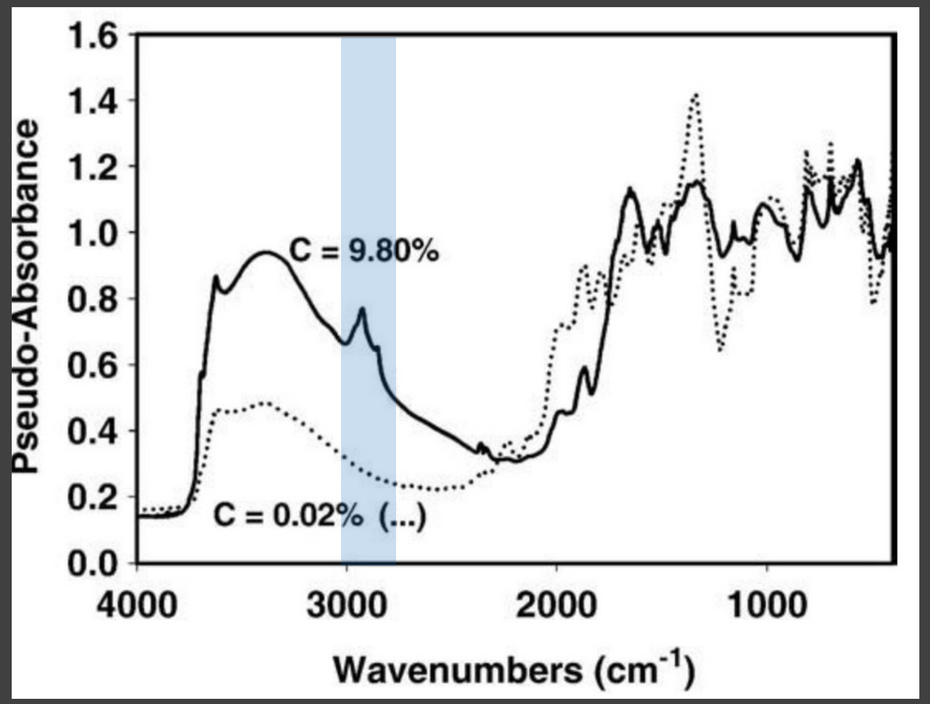
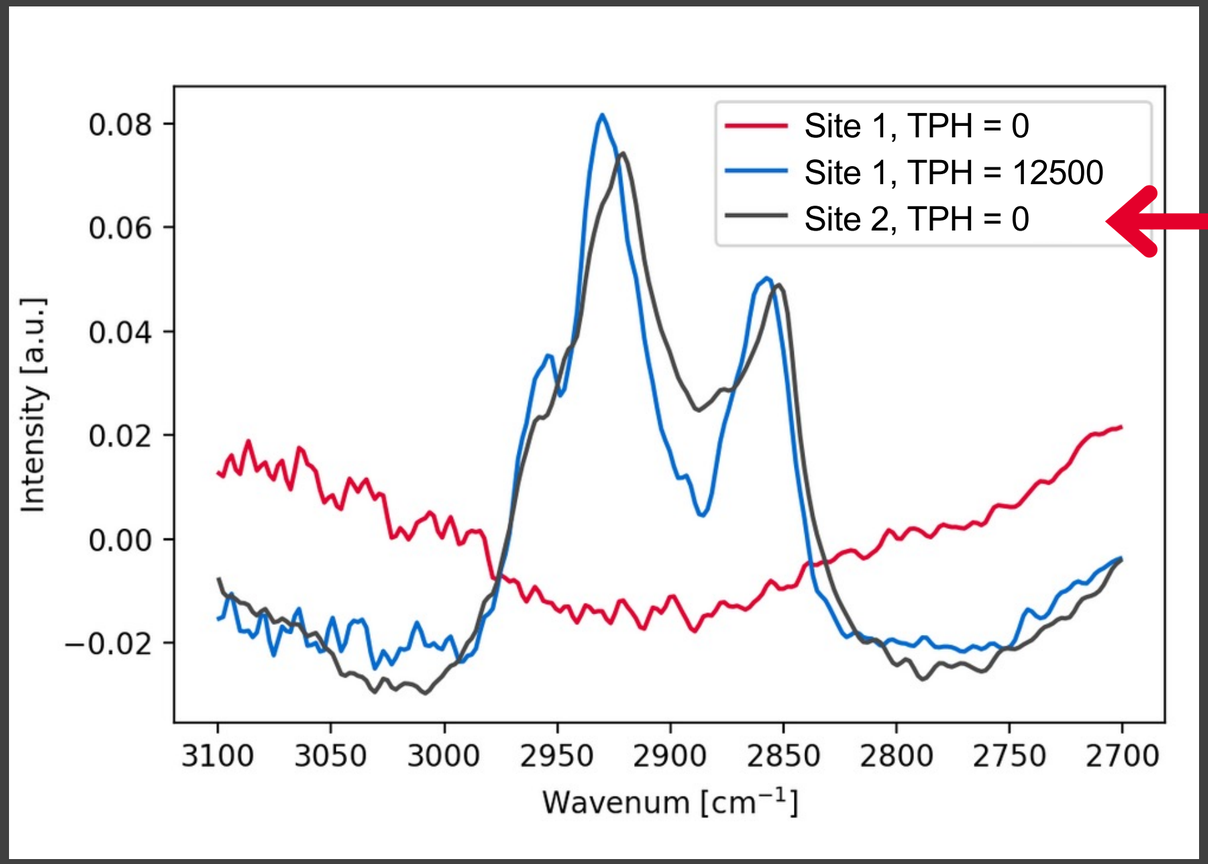


FIG: Mid-infrared spectra of high and low organic carbon (C) soils.

Reference: J.B. Reeves III., Near-versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasising carbon and laboratory versus on-site analysis: Where are we and what we needs to be done? DOI: 10.1016/j.geoderma.2009.04.005

With SOC



Ziltek

Edge case analysis

Calcium Carbonate [$CaCO_3$]

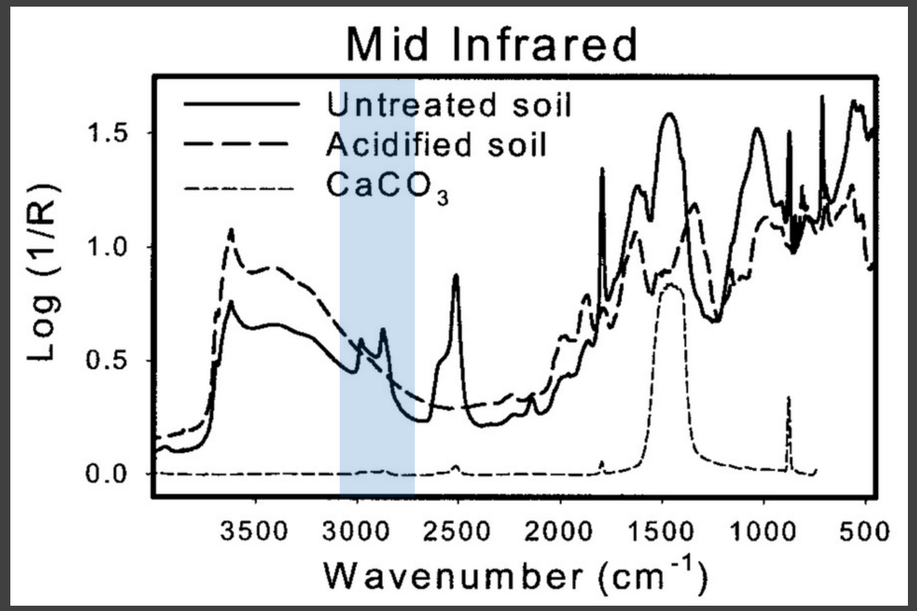
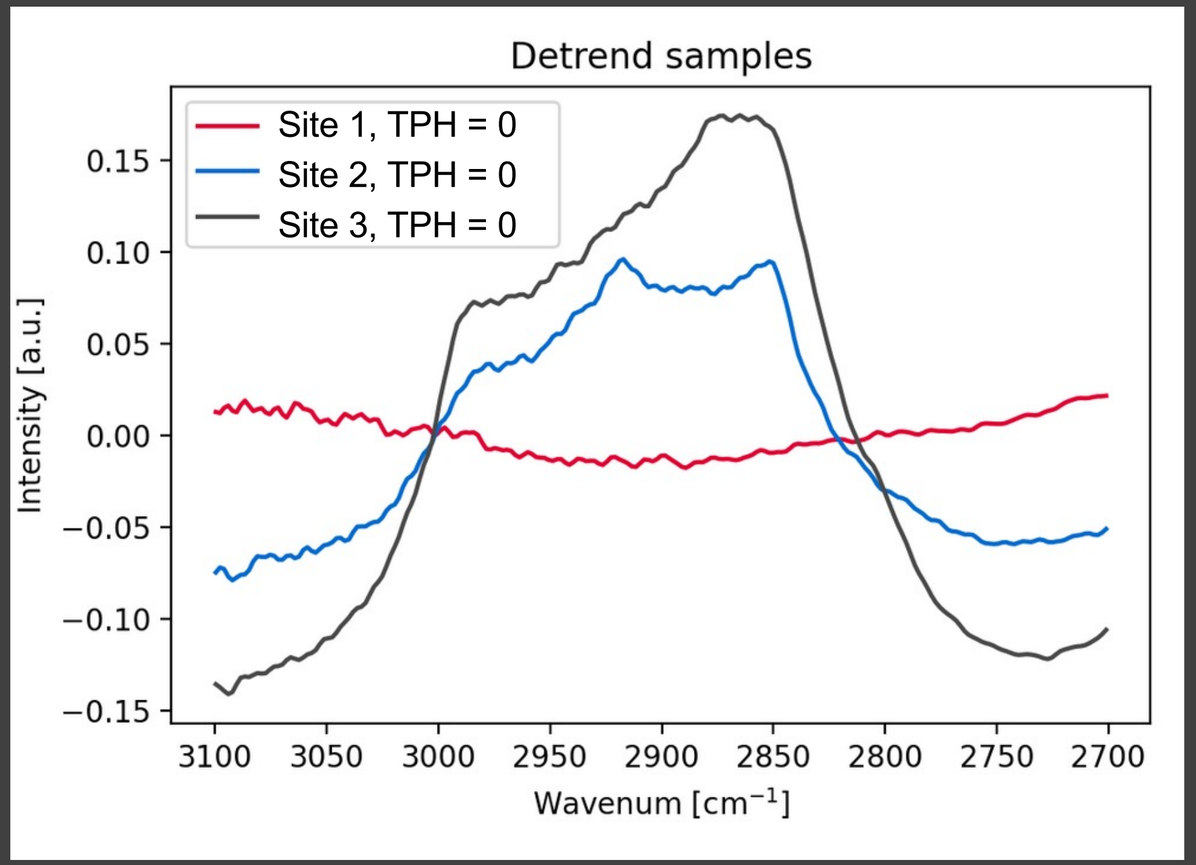


FIG: Comparison of mid-infrared and near-infrared spectra of a highly calcareous soil before and after treatment with acid for removal of carbonates. The carbonate (i.e., CaCO3) spectrum is included for additional comparison.

Reference: G. McCarty et al., Mid-Infrared and Near-Infrared Diffuse reflectance spectroscopy for Soil Carbon Measurements. DOI: 10.2136/sssaj2002.6400



Ziltek

Edge case analysis

Calcium Carbonate [$CaCO_3$]

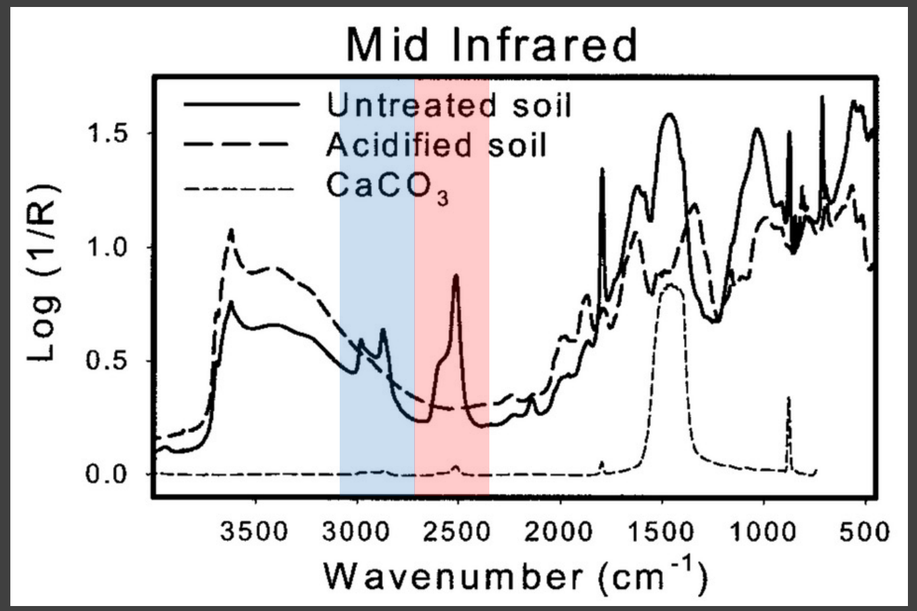
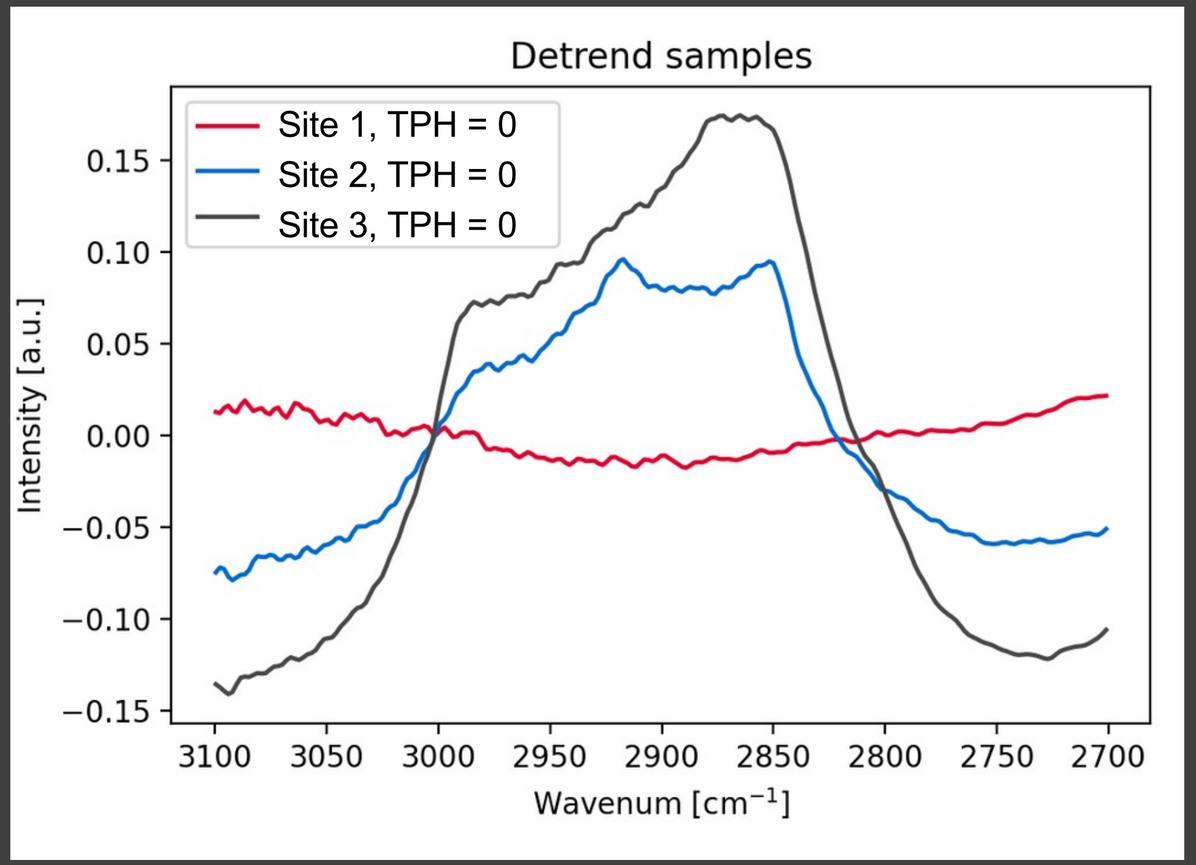


FIG: Comparison of mid-infrared and near-infrared spectra of a highly calcareous soil before and after treatment with acid for removal of carbonates. The carbonate (i.e., CaCO3) spectrum is included for additional comparison.

Reference: G. McCarty et al., Mid-Infrared and Near-Infrared Diffuse reflectance spectroscopy for Soil Carbon Measurements. DOI: 10.2136/sssaj2002.6400



Ziltek

Results

UPDATED TRAINING MODEL

- Samples removed: 2,824
  - High carbonate
  - Prominent SOC signatures

- Samples remaining: 11,751

Ensemble Majority Vote

	Predicted	Assigned
Sample 1-1	A	A
Sample 1-2	A	A
Sample 1-3	A	A
Sample 1-4	B	A
Sample 1-5	B	A

SVM Model Updates

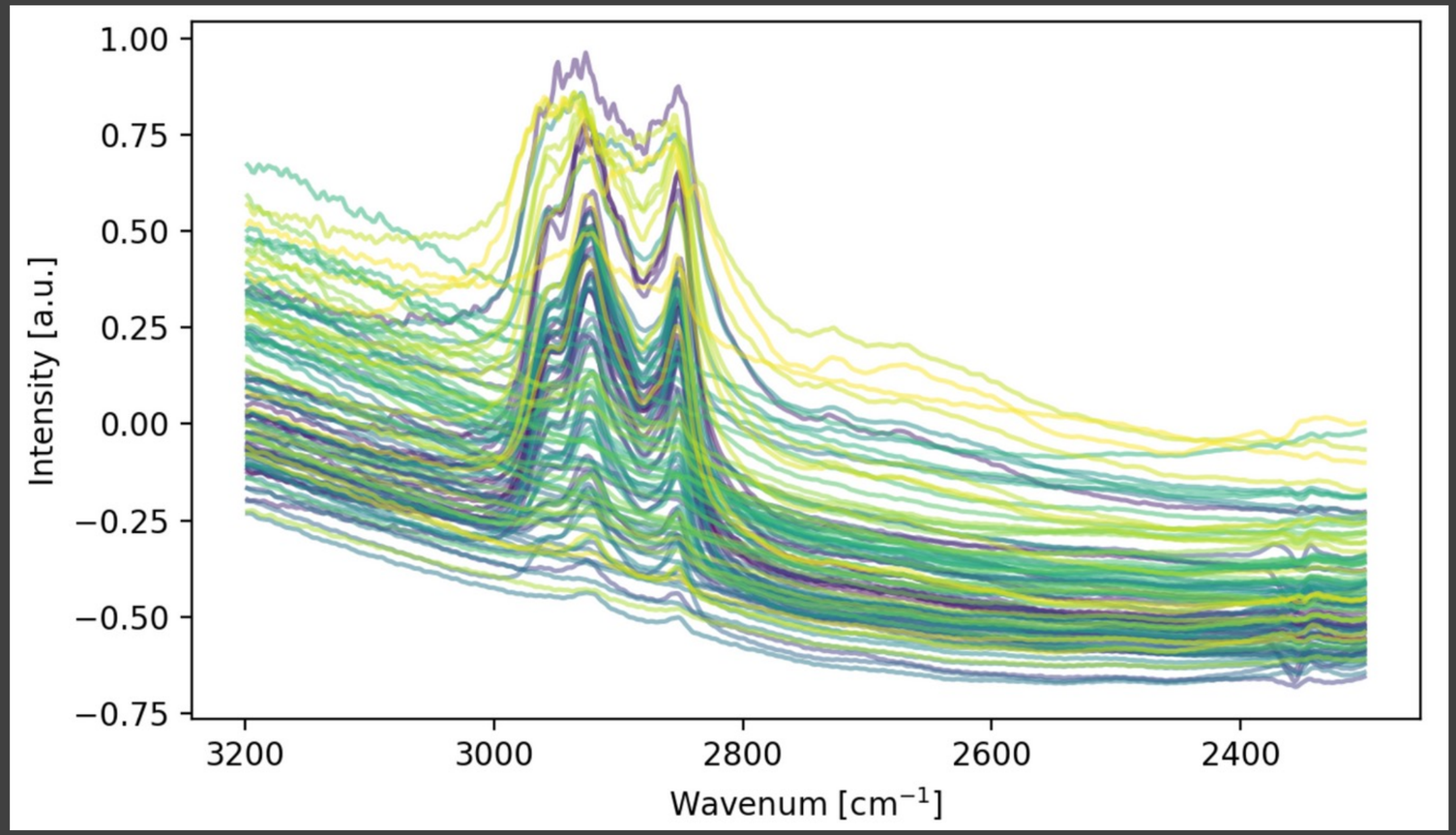
Complete Dataset	Filtered Dataset	Filtered data + Majority Voting
89%	90%	91%

Ziltek

Results

TEST CASE 1 - Indonesia

- Project Details
  - Sumatra
  - Contaminant: Crude Oil
  - Large scale, multi-year project
- Number of samples: 2038
  - Class A: 1228 (TPH > 1000)
  - Class B: 810 (TPH < 1000)
- Binary classifier: SVM
  - Accuracy: 95%
  - F1 Score: 0.931
  - MCC: 0.89

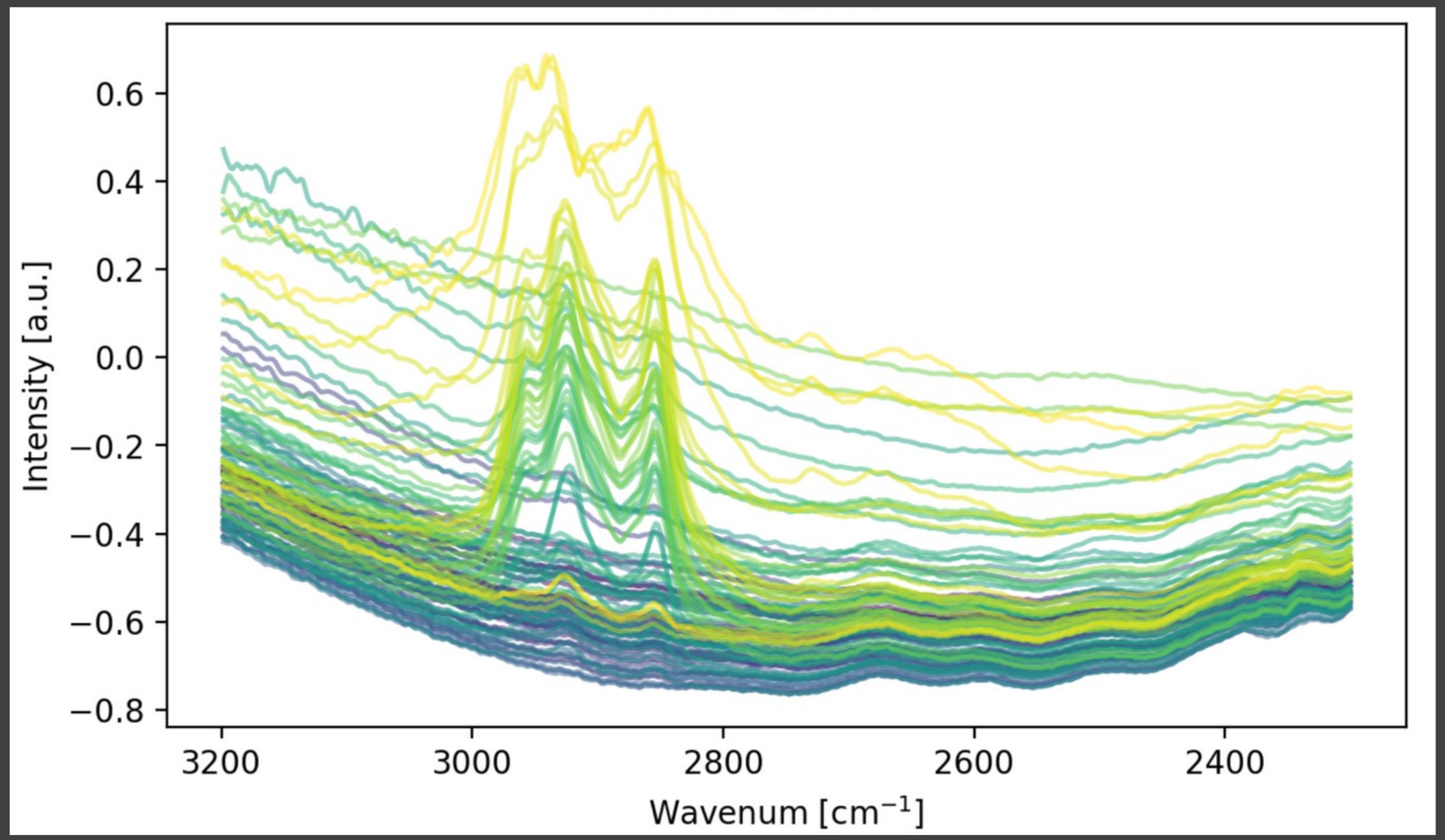


Ziltek

Results

TEST CASE 2 - Coastal Victoria

- Project Details
  - Coastal Wilderness
  - Contaminant: Crude Oil
  - Medium scale, 18 months project
- Number of samples: 553
  - Class A: 187 (TPH > 1000)
  - Class B: 366 (TPH < 1000)
- Binary classifier: SVM
  - Accuracy: 97%
  - F1 Score: 0.98
  - MCC: 0.94

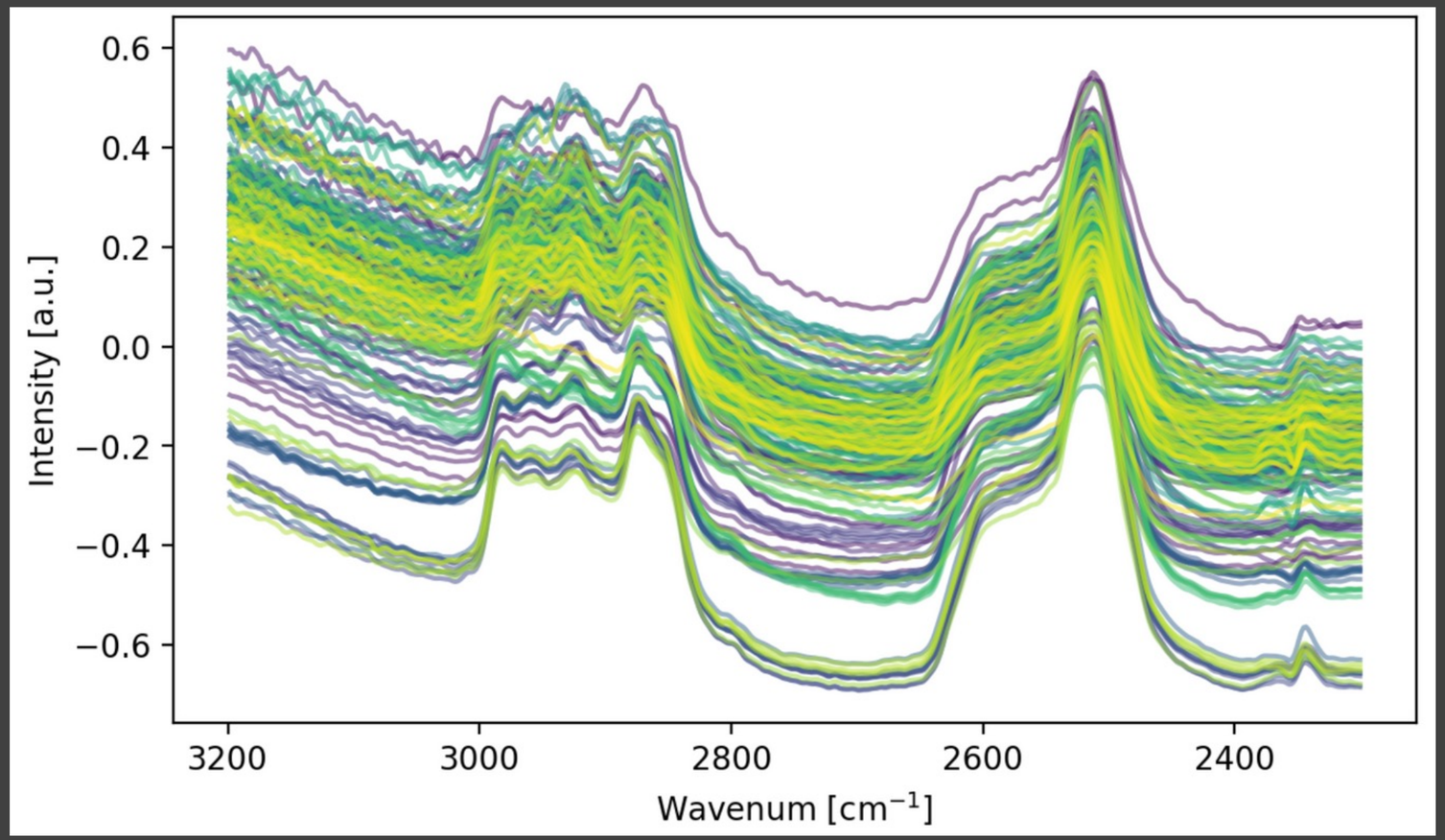


Ziltek

Results

TEST CASE 3 - France

- Project Details:
  - Industrial site
  - Contaminant: Diesel
  - Small scale, several weeks
- Number of samples: 612
  - Class A: 312 (TPH > 1000)
  - Class B: 300 (TPH < 1000)
- Binary classifier: SVM
  - Accuracy: 51%
  - F1 Score: 0
  - MCC: 0



HIGH IN CALCIUM CARBONATE !!!

Ziltek

Conclusion

- Developing binary classifier for rapid-assessment of hydrocarbon-contaminated soils.
- Tested 7 classifier models with 3 performance metrics.
- Training model accuracy: 90-99%
- SVM emerges as best model (Training & Testing dataset).
- Areas of further development
  - Edge cases (SOC, Carbonate)
  - Misclassified samples around threshold
- Historical data used to test performance in real world scenarios.
- Show promising results with accuracy ~ 90%.

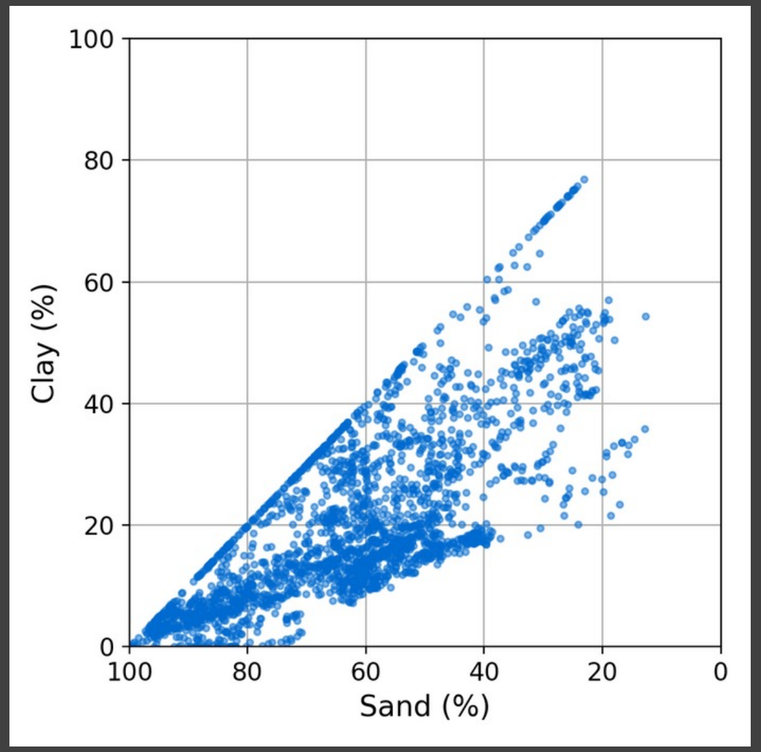
QUESTIONS?

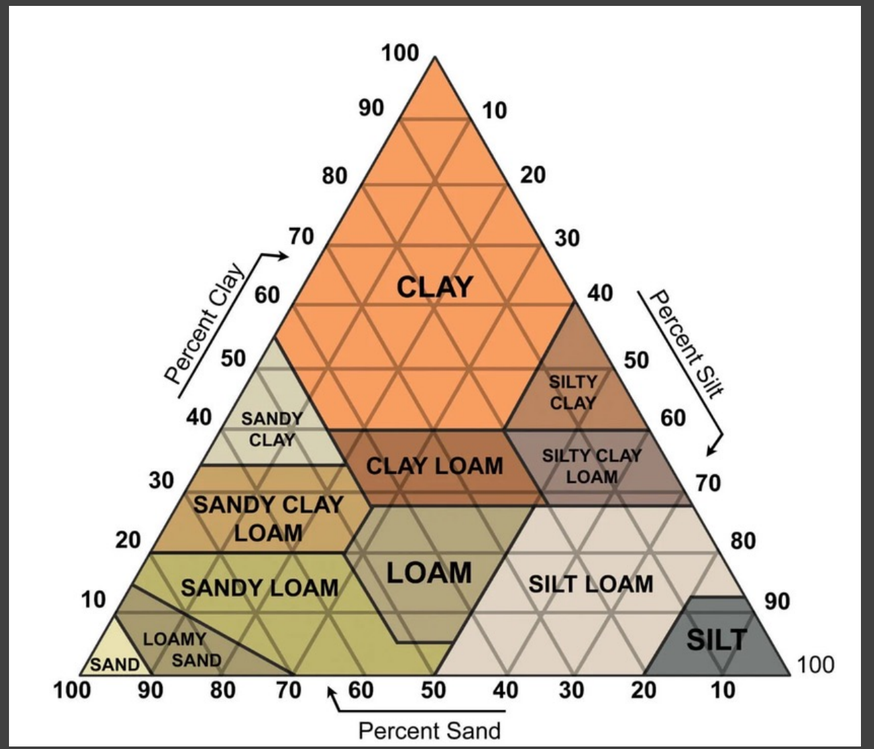
For further info, contact
info@ziltek.com

Ziltek

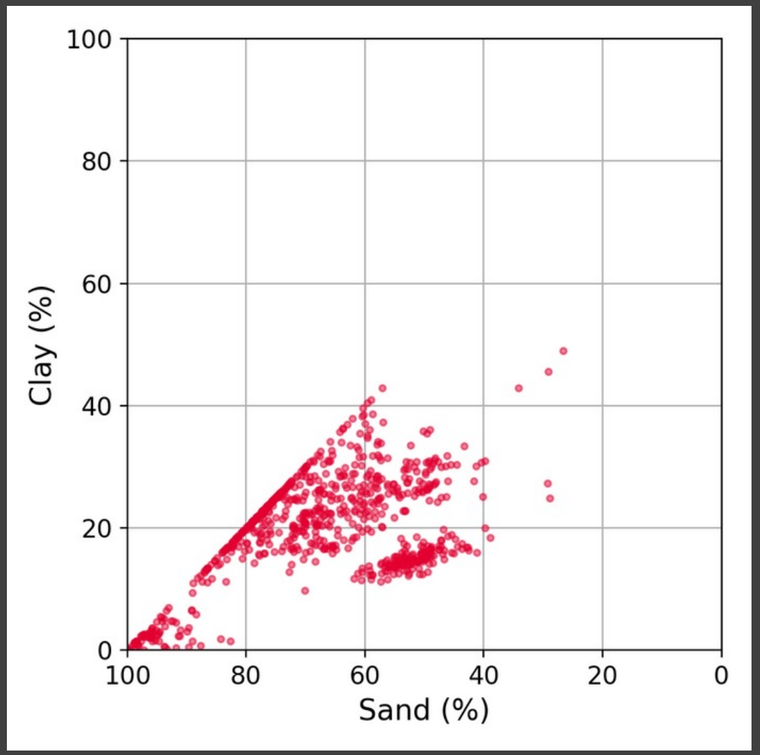
Soil Texture

Training Dataset





Testing Dataset



Ziltek

Confusion Matrix

Performance Metrics

ACTUAL	PREDICTED: POSITIVE	PREDICTED: NEGATIVE
POSITIVE	TP	FN
NEGATIVE	FP	TN

Positive = CONTAMINATED
Negative = CLEAN

TP= True Positive
FP = False Positive
FN = False Negative
TN = True Negative